APPLIED STATISTICAL REPORT

ON

ANALYSIS OF CAR PRICES IN UK

BY

PRAISE EKEOPARA

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The determination of car prices has been known to be dependent on some factors of which when not properly analyzed may become a challenge for futuristic intentions of reselling cars. Data collected from cars sold in UK were used as a case study, addressing this issues.

To effectively conduct these analysis, statistical methodologies such as exploratory data analysis, correlation analysis, hypothesis testing and predictive analytics were implemented to answer certain basic statistical questions. These questions include; knowing the current price trends, understanding features affecting the prices of cars in UK, testing significant differences between car prices and predicting accurately car prices.

The results obtained clearly shows that there was **99% increase** in the prices of cars from their start year to the year, 2019. However, there was a sharp fall of more than **70%** in the prices of cars in UK. It was equally observed that there were significant differences in the actual price of cars in UK. Also Engine size, mileage and year greatly affected the prediction of car prices.

INTRODUCTION

The prices of new cars when they are to be purchased are easily known from the sellers, but this can become a major challenge for an attempt to resale old or fairly used cars. This is as a result of unanalyzed considerable factors about the market before such sales are made.

This happens to be a case of an individual who currently finds it difficult to know the right price to sale off his old car as can be cited here. Hence immensely in need of the right insights and car price sale's model so as to make the best of decisions.

This report attempts to analyze such factors by developing and answering some statistical research questions. These answers will definitely prove useful for those interested in gaining insights of trends in car sales and trying to sale old cars just like in the aforementioned case.

The developed statistical related questions include the following;

1. what are the trends in car prices throughout the years?
2. which car brands have the highest and lowest prices in the market?
3. Is car mileage, mpg, engineSize, and tax good indicators of car prices?
4. which features strongly affects the outcome of car prices?
5. does the prices of some car's models significantly different from other models?
6. can we predict car prices based on the available features?

METHODOLOGY

In a quest to clearly answer the above outlined questions, the following suitable statistical methods were employed;

1. **Exploratory data analysis**: "In statistics, exploratory data analysis is an approach of analyzing datasets to summarize their main characteristics, often using statistical graphics and other data visualization methods" (Wikipedia, 2021). This method was specifically used to find useful insights such as trends in data.
2. **Correlation Analysis**: This is a process of finding the relationship existing between features in the given dataset. The use of visuals, Pearson, Spearman and other matrices can be used for this analysis, however Pearson's correlation was specifically used because it was able to measure the linear relationship between the present numerical features present in the data and the target label (the car price).
3. **Hypothesis testing**: This is a statistical method of calculating some quantities based on specific assumptions. The calculated results then will determine if the assumptions made will be accepted or rejected. There are various types of hypothesis testing techniques as illustrated in the Figure 1.0 below.

Variable Distribution type tests (Gussian)	Variable Relationship tests (Correlation)	Compare Sample Means (Parametric)	Compare Sample means (nonparametric)
• Shapiro-Wilk • D'Agostino's K^2 • Anderson-Darling	• Pearson's Correlation Coeff. • Spearman's Rank Coeff • Kendall's Rank • Chi-Squared test	• Student's t-test • Paired student's t-test • Analysis of variance (ANOVA) • Repeated Measures ANOVA test	• Mann-Whitney U test • Wilcoxon Signed-Rank • Kruskal-Wallis H test • Friedman Test

Figure 1.0: Types of Hypothesis testing techniques

In order as to give answer to the statistical question *"does the prices of some car's models significantly different from other models?"*, we need the help of a hypothesis test. We selected the parametric hypothesis (Paired t-test and the ANOVA) because we already know the mean distribution of the parameters to be tested. ANOVA and the Paired student's t-test measure the significance difference between the mean of groups. While the

pair t-test measures significance difference between mean of just two groups, it was used alongside the ANOVA test to validate our results. More details on these comparisons can be seen as shown fully in the result and discussion section.

4. **Predictive Analytics**: This is a type of data analytics that is employed whenever a statistical question involves a probabilistic answer. It was used to predict and determine what the car price can be, given a set of input parameters (obviously the car associated parameters). Four statistical regression models were built which include; Linear regression, Decision tree regression, k-nearest neighbor and the Random forest regression model. These regression models were used because our target variable happens to be a continuous data.

DATASETS

For the purpose of this analysis, scraped data of about 100,000 car listings were collected, which were separated into files corresponding to each car manufacturer. The data comprised of information on car prices and other useful information as described in Table 1.0 below.

Features	Description
1. model	The various car models for each car manufacturer such as A1, A2 etc.
2. year	Year for which car was purchased at a given price
3. price	The price of car purchased
4. transmission	Indicates type of car transmission e.g. automatic, manual, etc.
5. mileage	Measures car total distance traveled ever since it was manufactured.
6. fuelType	States the fuel type e.g. Petrol, Diesel, Hybrid etc.
7. tax	Defines the amount of tax placed on a car
8. mpg	Defines the miles per gallon rate of the car
9. engineSize	Defines the engine size
10. brand	This generally defines the car manufacturer e.g. bmw, focus, vw, mer etc.

Table 1.0: Data Features and descriptions

Although, the data collected was for cars sales in UK, it contains the necessary information such as the car prices, year, mileage, brand etc. and hence suitable for answering the aforementioned statistical questions.

DATA PREPROCESSING

In data analysis, data are usually prepared before they can be used for analysis this is because data collected maybe unclean by some data issues such as missing values, outliers etc. For the data available to us, the data preprocessing stages that was employed can be seen as summarized in the Figure 1.1 below.

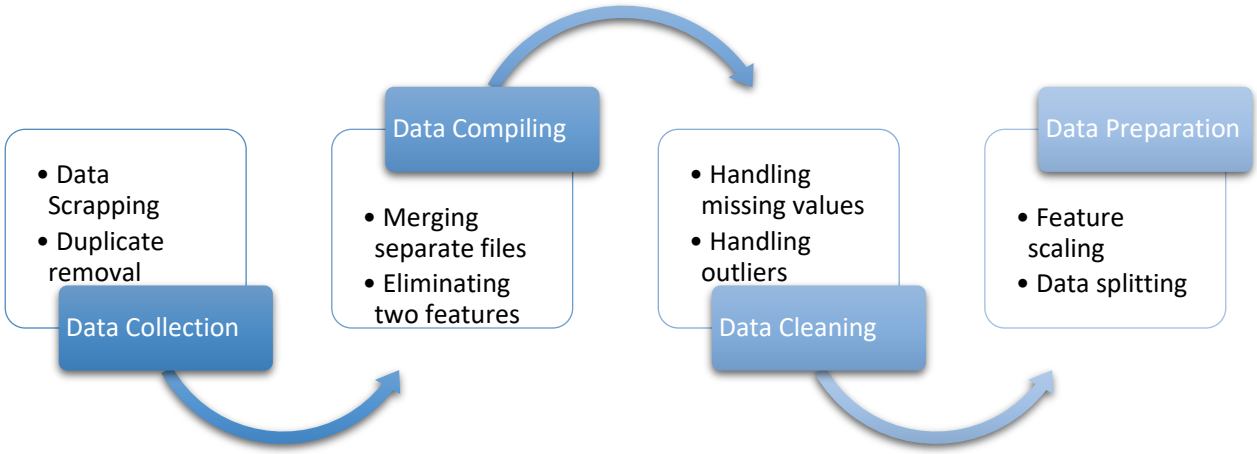

Figure 1.1: Data Preprocessing stages

- **Data Compilation:** When the data was collected, it was separated into files corresponding to each car manufacturer such as bmw, focus, merc, etc. hence it was imperative to compile all the separated files into a single data file so that all analysis can be carried out simultaneously. However, for the initial analysis, the two data files for **focus** and **cclass** were not included considering that they did not share same number of features with others.
- **Data Cleaning**: During the course of the analysis, certain data quality issues were encountered amongst which include; missing values, outliers and these were handled efficiently using the elimination method.
- **Data preparation**: The data was further prepared in algorithm consumable form. This include feature scaling of the data which is carried out so as to avoid the data being swamped with large absolute values as this will cause our algorithms to be biased when learning from the data and also the splitting of the data for training and testing of our model.

RESULTS AND DISCUSSION

Here, we discuss the findings from the statistical analysis conducted on the data using suitable visuals and tables.

a. Descriptive data analysis

The car prices as can be observed for each car manufacturer (brand) in Figure 1.2, which shows that the car prices for all brands generally experienced a hike in amount with the peak being in the year 2019.

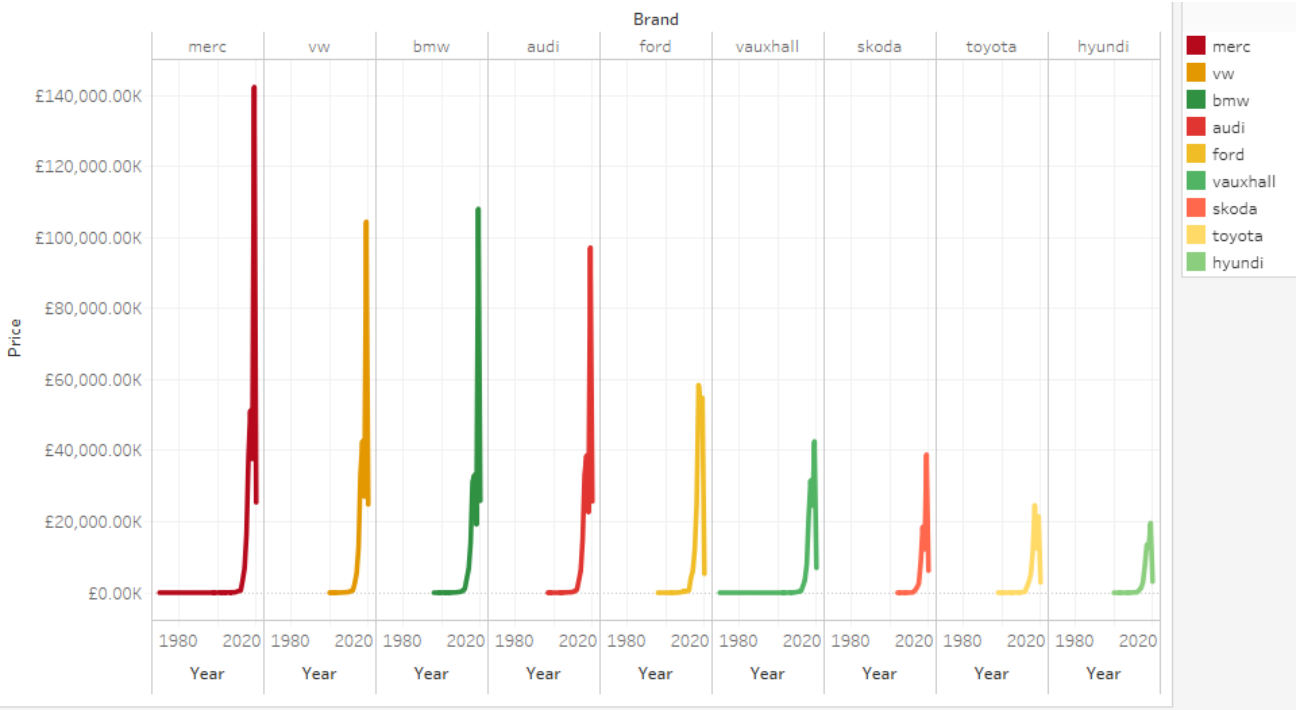

Figure 1.2: Car Prices by Brand and Year

Furthermore, from the analysis conducted, it was observed that there was more than 99% increase in the prices of cars from their start year to the year, 2019. However, a sharp fall of more than 70% decrease in the prices of cars in UK, this can be traced back to the effect of the pandemic within that time.

Figure 1.3 shows the various sum of car prices by car brands. It can be seen that the **Mercedes (merc)** brand happens to be costlier than others with about **£324,020.89** with the **Hyundi** being the cheapest at **£61,965.64.** However, the analysis carried out shows that the costliest car models for these car brands are Merc CClass and Hyundi Tucson.

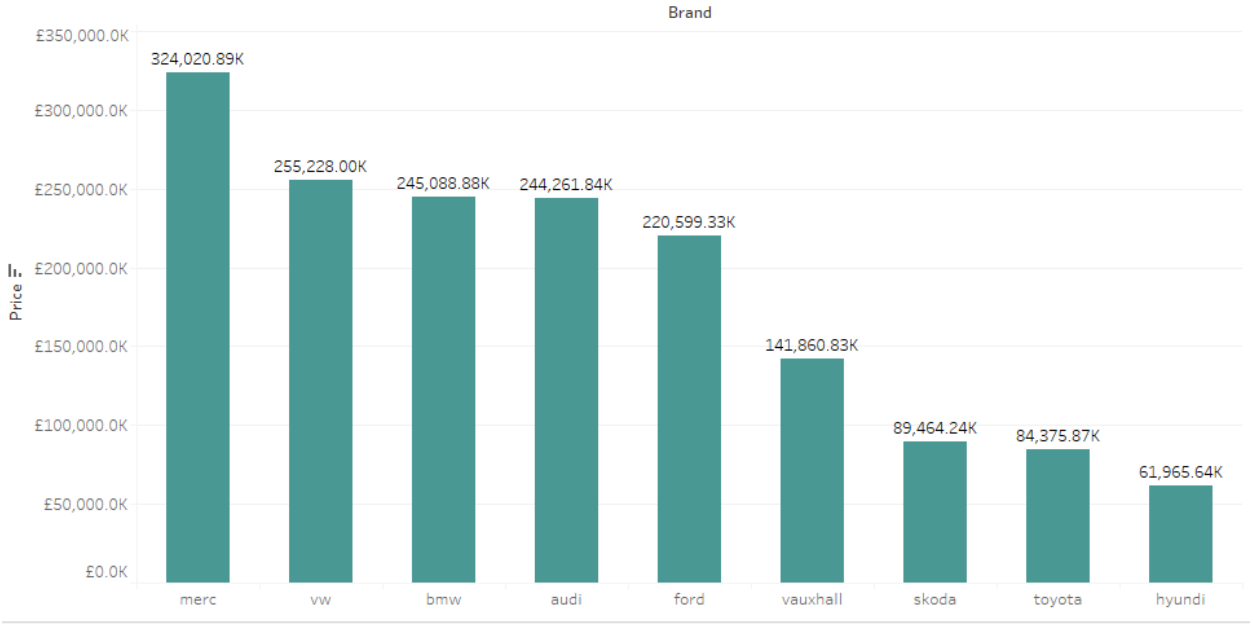

Figure 1.3: Car Prices by Brand

Figure 1.4, and 1.5 below illustrate car prices against transmission and fuel types being used. It can be observed that for transmission, the Semi-Auto happens to be slightly costlier than other transmission types. This can be attributed to its flexibility in choice of selection.

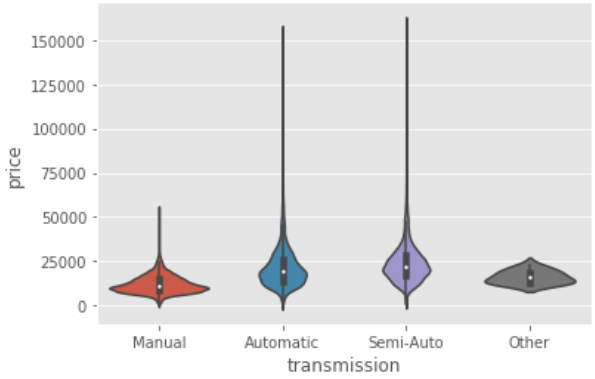

Figure 1.4: Prices by transmission

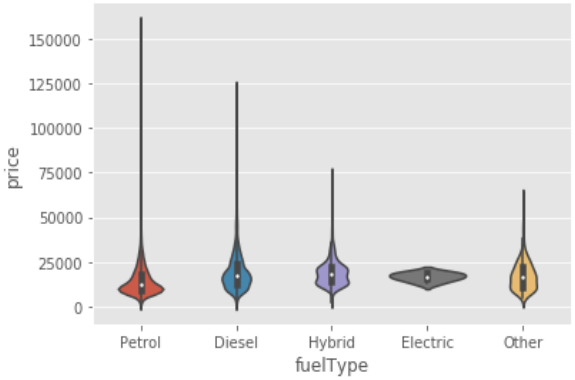

Figure 1.5: Car Prices by fuelType

However, it can be seen for the case of fuel types, that the electric cars happen to have the highest average of price. This can equally be attributed to possibly high demand that will be placed on it based on its durability, fuel cost savings, environmental friendliness and other factors.

b. Correction Analysis

Here the Pearson's correlation was used to determine the relationship between car features and the car prices (the target feature). As the coefficient of determination approaching 1 which depicts stronger relation progression, the various scores for each feature and the other are therefore summarized in Table 1.1 below.

	year	price	mileage	tax	mpg	engineSize
year	1.000000	0.492059	-0.740443	0.195825	-0.132909	-0.038560
price	0.492059	1.000000	-0.417946	0.307667	-0.296440	0.638113
mileage	-0.740443	-0.417946	1.000000	-0.220786	0.185671	0.109081
tax	0.195825	0.307667	-0.220786	1.000000	-0.451446	0.278420
mpg	-0.132909	-0.296440	0.185671	-0.451446	1.000000	-0.248214

Table 1.1: Correction table for price prediction

It can be observed from the heatmap plot in Figure1.6 below, that EngineSize happens to be a good predictor of car prices with score of 0.64. The year and tax exhibited weak correlations of 0.49 and 0.31 respectively while other features exhibited very weak and negative relationship.

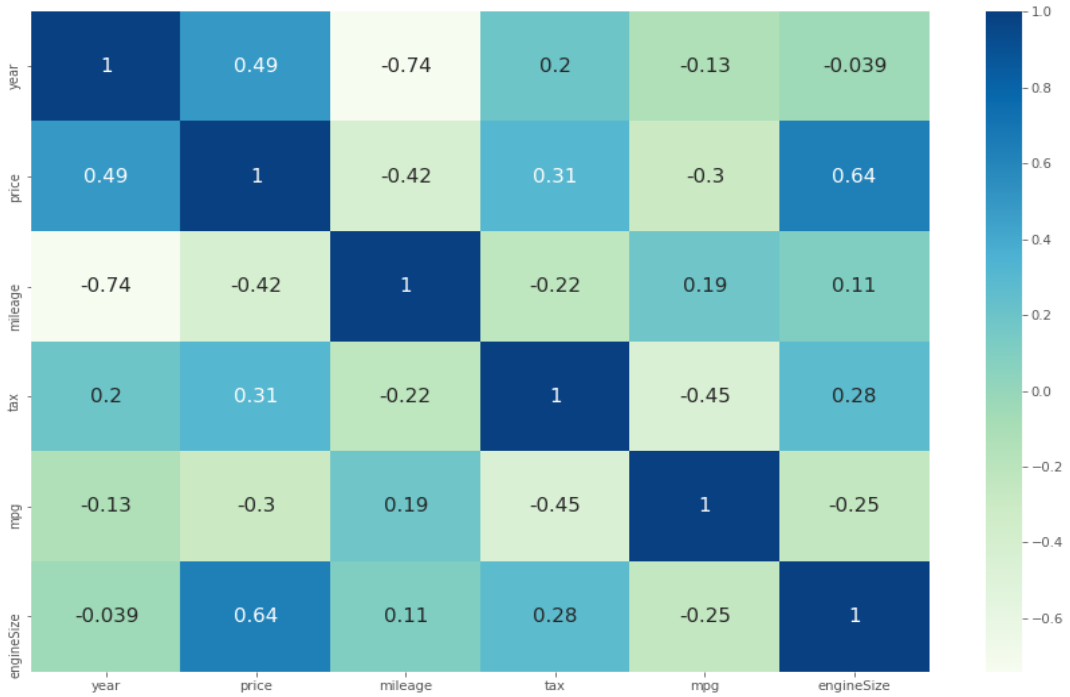

Figure 1.6: Heatmap Plot of Car features

c. Hypothesis Testing

The two parametric hypotheses (Paired t-test and the ANOVA) being used for testing the significance difference between the average car prices were used because of the following assumptions that were satisfied, which include;

- Population distribution is normal, and
- Samples are random and independent
- The sample size is small.
- Population standard deviation is not known.
- Homogeneity of sample variance

Most of these assumptions are shown in the boxplots of Figure 1.7 below, which shows the various distributions of the car prices from each brand.

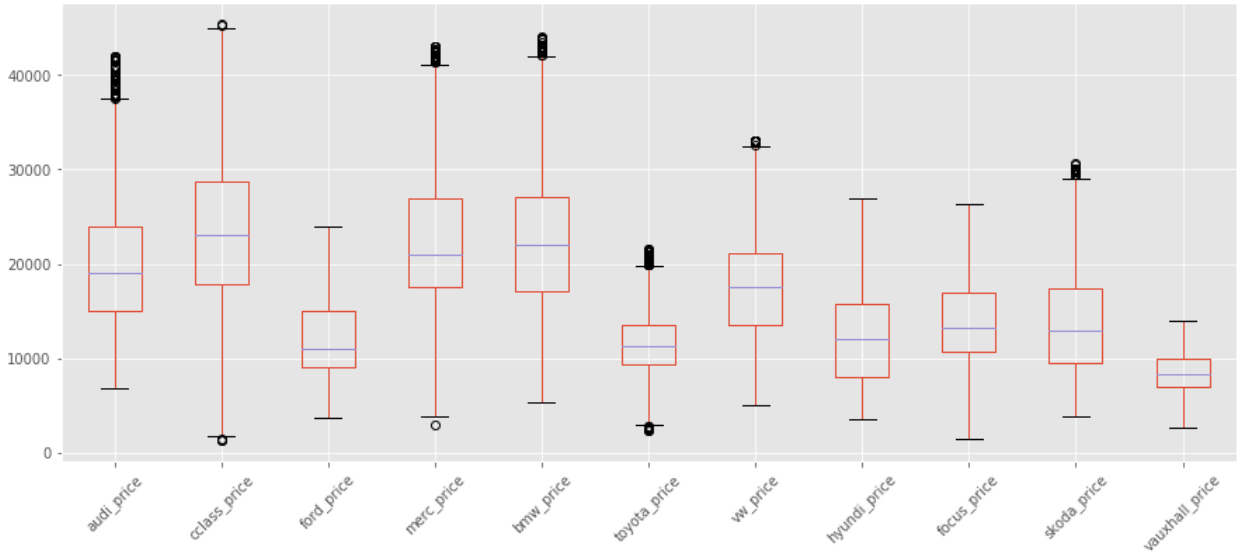

Figure 1.7: Car price distributions for each brand

It can be observed from the boxplot above, that there happens to be an average significance difference between the various car prices. However, we cannot conclude from this, hence the parameters were put to test.

After subjecting the parameters to test, the both results can be seen in Table 1.3 below, it was seen that the p-value was 0.00 and hence less than any significance interval that could

be used, therefore the null hypothesis was rejected and we concluded by inferring that the average prices of all car prices were unequal.

Hypothesis test	p-value	Statistics (stat)	Conclusion
ANOVA (f_oneway)	0.000	4304.436	The average prices for all brands are unequal
Paired t-test (ttest_rel)	0.000	47.458	The average prices for all brands are unequal

Table 1.3: Hypothesis testing results and conclusion

d. Predictive Analysis

In this section, four models were built for prediction of car prices when inputted a set of input variables. The models and their various accuracy scores based on the test scores can be seen in the Table 1.4 below.

Models	Accuracy Score
1. Linear Regression	0.717698
2. Decision Tree	0.885853
3. K-Nearest Neighbor	0.920235
4. Random Forest Regression	0.940200

Table 1.4: Models and their accuracy scores

The Random forest regression model outperformed others with an accuracy score of 93.96% and hence was selected to be the best performing model and recommended for car price prediction.

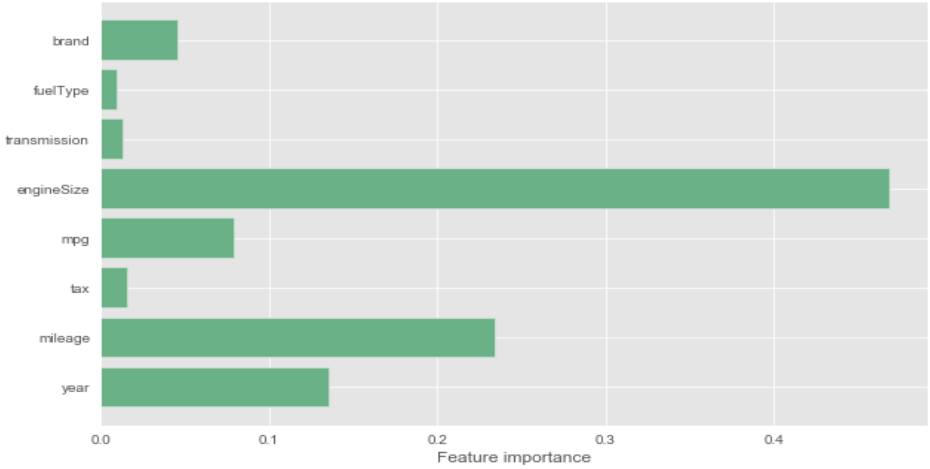

Figure 1.8: Feature importance

Figure 1.8 therefore shows the feature importance for predicting the car prices using the Random Forest regression model. It can be seen that engineSize, Mileage and year were considered as most important features affecting price. This equally confirms the high correlation scores by these three features as was discovered in the correlation analysis section.

CONCLUSION

The analysis on car sales in the United Kingdom has been successfully carried out using the collected dataset and hence providing suitable answers to the statistical questions asked.

The summary of the key findings from this analysis are then as follows;

- The **year 2019** recorded the highest price of cars in UK with more than **99% increase** in the prices of cars from their start year to the year, 2019. However, a there was a sharp fall of more than **70% decrease** in the prices of cars in UK, which can be traced back to the effect of the pandemic within that time.
- Mercedes brand happens to be costlier than others with about **£324,020.89** with the Hyundi being the cheapest at **£61,965.64.** However, the analysis carried out shows that the costliest car models for these car brands are **Merc CClass** and **Hyundi Tucson**.
- **Semi-Auto** cars happens to be slightly costlier than other transmission types while the **electric cars** happen to have the highest average of price. This can be attributed to their flexibility in transmission choice and reduction in costs respectively.
- The prices of some car's models are significantly different from other models
- Determination of car prices are largely dependent on the **engine size**, the **year** and the **mileage** rate of the car.
- The **Random forest regression** model outperformed others with an accuracy score of **94.02%** and hence was selected to be the best performing model and recommended for car price prediction.

References

Aashi, G., 2021. *Analytics Vidhya.* [Online]
Available at: https://www.analyticsvidhya.com/blog/2021/06/hypothesis-testing-parametric-and-non-parametric-tests-in-statistics/
[Accessed 1 June 2021].

Aditya, 2020. *Kaggle.* [Online]
Available at: https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes
[Accessed 2021].

Jason, B., 2020. *Navigation.* [Online]
Available at: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/
[Accessed 20 August 2021].

Jason, B., 2020. *Navigation.* [Online]
Available at: https://machinelearningmastery.com/statistical-hypothesis-tests/
[Accessed 10 April 2021].

Jason, B., 2020. *Navigation.* [Online]
Available at: https://machinelearningmastery.com/feature-selection-with-categorical-data/
[Accessed 18 August 2021].

Scipy, 2008. *API Reference.* [Online]
Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html
[Accessed 2021].

University, A., n.d. *Pytolearn.* [Online]
Available at: http://pytolearn.csd.auth.gr/d1-hyptest/12/anova-one.html

Wikipedia, 2021. [Online]
Available at: https://en.wikipedia.org/wiki/Exploratory_data_analysis